AMERICAN ACADEMY OF OPHTHALMOLOGY®

Artificial Intelligence Models to Identify Patients with High Probability of Glaucoma Using Electronic Health Records

Rohith Ravindranath, MS, Sophia Y. Wang, MD, MS

Purpose: Early detection of glaucoma allows for timely treatment to prevent severe vision loss, but screening requires resource-intensive examinations and imaging, which are challenging for large-scale implementation and evaluation. The purpose of this study was to develop artificial intelligence models that can utilize the wealth of data stored in electronic health records (EHRs) to identify patients who have high probability of developing glaucoma, without the use of any dedicated ophthalmic imaging or clinical data.

Design: Cohort study.

Participants: A total of 64 735 participants who were ≥18 years of age and had ≥2 separate encounters with eye-related diagnoses recorded in their EHR records in the All of Us Research Program, a national multicenter cohort of patients contributing EHR and survey data, and who were enrolled from May 1, 2018, to July 1, 2022.

Methods: We developed models to predict which patients had a diagnosis of glaucoma, using the following machine learning approaches: (1) penalized logistic regression, (2) XGBoost, and (3) a deep learning architecture that included a 1-dimensional convolutional neural network (1D-CNN) and stacked autoencoders. Model input features included demographics and only the nonophthalmic lab results, measurements, medications, and diagnoses available from structured EHR data.

Main Outcome Measures: Evaluation metrics included area under the receiver operating characteristic curve (AUROC).

Results: Of 64 735 patients, 7268 (11.22%) had a glaucoma diagnosis. Overall, AUROC ranged from 0.796 to 0.863. The 1D-CNN model achieved the highest performance with an AUROC score of 0.863 (95% confidence interval [CI], 0.862–0.864). Investigation of 1D-CNN model performance stratified by race/ethnicity showed that AUROC ranged from 0.825 to 0.869 by subpopulation, with the highest performance of 0.869 (95% CI, 0.868–0.870) among the non-Hispanic White subpopulation.

Conclusions: Machine and deep learning models were able to use the extensive systematic data within EHR to identify individuals with glaucoma, without the need for ophthalmic imaging or clinical data. These models could potentially automate identifying high-risk glaucoma patients in EHRs, aiding targeted screening referrals. Additional research is needed to investigate the impact of protected class characteristics such as race/ethnicity on model performance and fairness.

Financial Disclosure(s): The author(s) have no proprietary or commercial interest in any materials discussed in this article. *Ophthalmology Science 2025;5:100671* © *2024 by the American Academy of Ophthalmology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).*

Supplemental material available at www.ophthalmologyscience.org.

Glaucoma is a debilitating group of eye diseases that cause progressive damage to the optic nerve, leading to permanent vision loss and blindness if left untreated. It is the leading cause of blindness worldwide[1–3] and in the United States,[4] currently affecting 80 million people worldwide and 3 million people in the United States.[4] In its early stages, glaucoma is often asymptomatic, with vision loss progressing gradually over time. Up to 50% to 80% of glaucoma cases are undiagnosed.[5,6] In addition, glaucoma disproportionately affects Black and Hispanic/Latino populations.[7–9] Not only do these high-risk populations suffer from a higher burden of undetected and untreated glaucoma,[6] but they also have more advanced glaucoma by the time they are diagnosed,[10] are more likely to lose vision after glaucoma diagnosis,[10] and are underrepresented overall in glaucoma research.[11]

Glaucoma screening can facilitate early diagnosis and treatment of glaucoma; previous screening studies in local population-based samples have estimated the screen-positive rate to be approximately 1% to 2% in Whites,[12] 2% in Hispanics,[13] and up to 5% in Blacks.[12] However, the United States Preventive Services Task Force has identified several barriers to widespread screening.[14] Glaucoma diagnosis requires assessment of the structure

© 2024 by the American Academy of Ophthalmology
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/). Published by Elsevier Inc.

https://doi.org/10.1016/j.xops.2024.100671
ISSN 2666-9145/25

and function of the optic nerve. Most previous research has been carried out on glaucoma risk stratification with imaging or functional tests,[15–19] and "most tests require specialized equipment and are performed in an eye specialty setting,"[14] rendering widespread screening/diagnosis with these modalities less feasible. Furthermore, there is a lack of reliable tools for identifying individuals at a higher risk of developing glaucoma who would be the optimal candidates for resource-intensive glaucoma screening.[14] Research is necessary to develop and assess risk assessment tools, both in early detection of those at an elevated risk for glaucoma and in guiding effective screening approaches.[14]

Previous studies for identifying high-risk patients at the prescreening stage have generally been limited in scope, focusing on race/ethnicity and family history.[20–22] Without the ability to target equipment- and resource-intensive screening efforts to the high-risk populations, it has been difficult to conduct randomized controlled trials with the magnitude and follow-up required to definitively demonstrate the benefits of glaucoma screening on long-term vision-related quality of life, even though treatment of glaucoma is known to prevent vision loss over the lifetime of the patient.[14]

The National Institutes of Health's All of Us[23] Research Program presents a unique opportunity to develop glaucoma prescreening tools to identify patients with a high probability of glaucoma in a large and diverse population of patients in the United States, using artificial intelligence (AI) on electronic health record (EHR) data, thus addressing the above key barriers. The platform provides an extensive range of health information from over a quarter million individuals across the United States and self-reported survey data. The research program also makes rigorous efforts to capture an especially diverse population, which includes traditionally underrepresented groups in clinical research.[24] The purpose of this study was to develop AI prediction algorithms that can analyze the wealth of data stored in EHR on diagnoses, medications, and labs to identify patients who have high probability of glaucoma, without the use of any ophthalmic imaging or clinical data. Algorithms using this approach could one day be used as a prescreening tool to automatically identify patients with a higher probability of glaucoma, targeting the more resource-intensive image-based glaucoma screening programs to these people.

Methods

We conducted a retrospective cross-sectional study using the All of Us Research Program to develop comprehensive EHR predictive models that incorporate systemic diagnosis, medications, laboratory results, clinical information (such as blood pressure), and demographic data to predict the risk of glaucoma. This study using deidentified data was exempt from Stanford University Institutional Review Board approval and adheres to the tenets of the Declaration of Helsinki.

Data Source

The All of Us Research Program has established a longitudinal database that includes clinical, environmental, lifestyle, and genetic data from >250 000 individuals in the United States. Participation in the program is open to individuals aged ≥18 years of age residing in the United States, who can join through the program's website (JoinAllOfUs.org) or >60 health care provider organizations. Participants contribute data from their EHRs and fill out surveys reporting demographic and other information. All data are mapped to the Observational Medical Outcomes Partnership Common Data Model for observational health data.[25]

Inclusion/Exclusion Criteria

This study included patients from the All of Us, version 7 cohort, the latest data release as of the time of analysis, which included data from participants enrolled from May 1, 2018, to July 1, 2022. Participants had ≥2 separate encounters with eye-related diagnoses in their EHR records and therefore are likely to have visited an eye care provider and had the opportunity for glaucoma status to be assessed. As we wished to train a model to identify patients who had definitive glaucoma, patients with definitive glaucoma diagnosis codes and patients without any glaucoma-related diagnosis codes were included; patients who were diagnosed with suspect or borderline glaucoma were excluded from the study (see Table S1, available at www.ophthalmologyscience.org, for related concept codes). Patients with no previous eye-related diagnoses at all were also excluded, as it was not possible to determine whether these patients might have had undiagnosed glaucoma. The cohort design and inclusion/exclusion criteria for the study population are summarized in Figure 1.

Measures

Input features from EHRs included demographics, systemic (nonocular) medications, systemic (nonocular) diagnoses, and lab results. For patients with a diagnosis of glaucoma, we only included data from the EHR that was recorded before the patient's initial glaucoma diagnosis. Demographics included age, sex at birth, race, and ethnicity. Age was a continuous variable. Gender, race, and ethnicity were categorical variables that were dummy encoded for model input. Lab results and measurement variables included body mass index (BMI), diastolic blood pressure, systolic blood pressure, heart rate, hemoglobin A1c, thyroid-stimulating hormone, lipid panel, complete blood count panel, comprehensive metabolic panel, and pH of urine. Lipid panel, complete blood count panel, and comprehensive metabolic panel all included multiple lab results as each panel has multiple components. All lab results and measurement variables were normalized. Missing values for demographics and lab/measurement results were imputed using k-Nearest Neighbors technique,[26] which has been used in previous studies predicting glaucoma.[26] All systemic (nonocular) medications and (nonocular) diagnoses were included as model input. We excluded all ocular medications (Observational Medical Outcomes Partnership Parent Codes: 21603551, 21605125, 21605126, 21605187), including oral carbonic anhydrase inhibitors, and we excluded ocular diagnoses (Observational Medical Outcomes Partnership Parent Codes: 118235002), leveraging the hierarchical nature of the ontology by excluding broad eye-related parent codes and all of their child concept codes as well. Only nonocular features were included, as the purpose of the model is to be able to be deployed on patients with potentially no prior eye examinations or encounters.

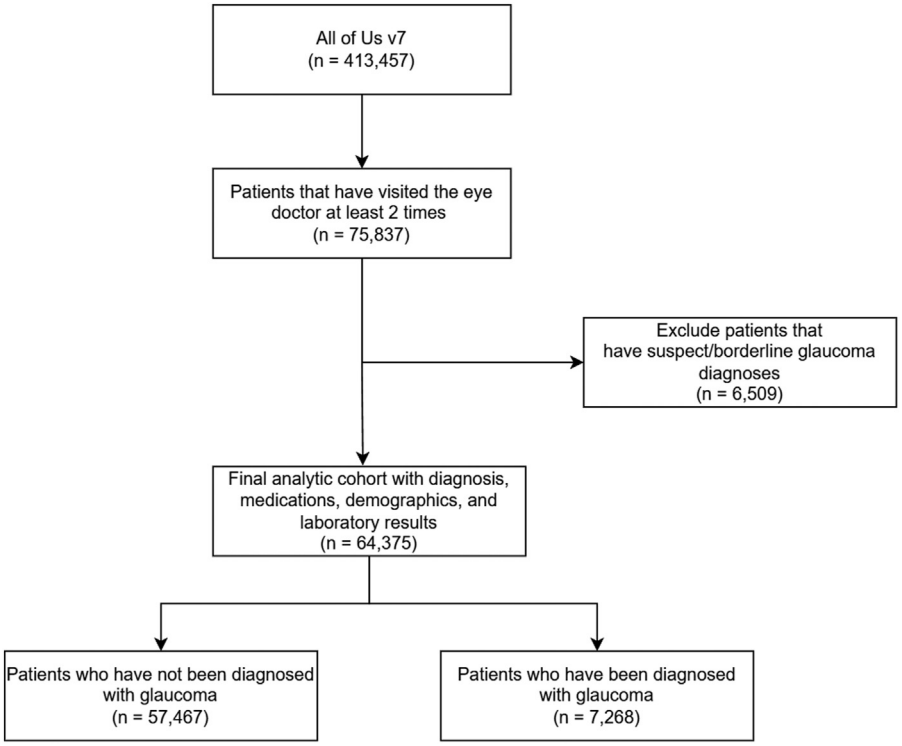

Figure 1. Feature engineering and cohort construction. Flowchart depicting the cohort design, with inclusion and exclusion criteria for the study population.

Diagnosis codes (only nonocular) included were converted to Boolean vectors such that patients with an encounter with that diagnosis code have a "1" and 0 otherwise. Diagnosis codes with near-zero variance (<0.005%) were removed, yielding a total of 4891 diagnosis code−based features. Similarly, medications (only nonocular) were turned into Boolean vector inputs as with diagnosis codes. Medication features with near-zero variance (<0.005%) were removed, yielding a total of 4618 medication features remaining. The total number of structured input features was 9553. A full list of features is included in Table S2 (available at www.ophthalmologyscience.org). The cohort was divided randomly into a training set of N = 51 788 and a test set of N = 12 947.

Modeling

Machine Learning. Machine learning models were fitted on the training data using the Python sklearn v1.2.1 package.[27] These models included penalized logistic regression (L2 penalization) and XGBoost. Hyperparameters were tuned using fivefold cross-validation on the training set to optimize the area under the receiver operating curve (AUROC). Grid search was used to tune penalization for logistic regression, and random search was used for XGBoost. A summary of hyperparameters is present in Table S3 (available at www.ophthalmologyscience.org).

Deep Learning. Two deep learning models were built and evaluated using the Python TensorFlow v2.11.0 package.[28] The first deep learning model was a fully connected network using the same structured features as the machine learning models, constructed with 3 dense layers, 2 dropout layers, and an output layer with a sigmoid activation function [Input(9,568) -> Dense(512) -> Dropout(0.4) -> Dense(256) -> Dense(128) -> Dropout(0.4) -> Output(1, sigmoid)].

The second deep learning model was a custom architecture consisting of 2 stacked autoencoders and a 1-dimensional convolutional neural network (1D-CNN) model. In the first phase, 2 stacked autoencoders were trained to learn encodings (latent size of 128 features) for diagnoses and medications separately. An autoencoder, an unsupervised learning algorithm, seeks to learn efficient representations of input data by reconstructing the original input as its output. This process is facilitated through a twofold structure comprising an encoder and a decoder.[29] The encoder transforms input data into a reduced dimensional representation, often termed as the "latent space" or "encoding." From that representation, a decoder rebuilds the initial input. A stacked autoencoder is a multilayer neural network that consists of multiple autoencoders, where the output of each encoder is fed into the next encoder until the last encoder feeds its output into a chain of decoders. This sequential arrangement facilitates a systematic compression and decompression of input data, affording greater control over each stage of the process. Stacked autoencoders are an effective method for dimensionality reduction, especially when capturing nonlinear and complex relationships in the data is essential. In this study, both stacked autoencoders had the same architecture, where the inputs are entered into a stack of 4 encoders: [input_dim -> 1028 -> 512 -> 256 -> 128]. The final encoded result is the required input to a stack of 4 decoders. The decoders attempt to reverse the previous encoding process: [256 -> 512 -> 1028-> output_dim]. During evaluation, only the encoders are used for reduced-dimensional representation. The second phase consists of the 1D-CNN. The inputs of the model consist of the outputs of the 2 encoders concatenated with the remaining input variables (demographics, lab results, and measurements), resulting in 300 features. Figure 2 illustrates the architecture of our 1D-CNN. The model begins by increasing the feature dimension through a fully

connected layer. The role of this layer is to soft-order the features in our data to create spatial locality and provide enough "pixels" for the subsequent convolutional layers. Data is then resized as 64@16 × 1. In simple words, each of these signals corresponds to a group of 16 ordered feature representations, and we have 64 groups with different orderings. Data are then fed through 3 convolutional layers, a max pooling layer, flattened, and finally passed through a sigmoid activation output layer. Figure 3 showcases the full prediction pipeline incorporating 1D-CNN and autoencoders.

Evaluation

We used standard classification evaluation metrics including accuracy, balanced accuracy, recall, precision, F1 score (the harmonic mean of recall and precision), AUROC, and the area under the precision−recall curve, all evaluated on the test set. Accuracy measures the percentage of correctly predicted labels, while balanced accuracy avoids inflation of accuracy metrics resulting from imbalanced data by taking the average between the sensitivity and the specificity, which is equivalent to the accuracy where each sample is weighted by the inverse prevalence of its true class. The classification threshold for each model was tuned during cross-validation for the optimum F1 score on the test set. Ninety-five percent confidence intervals (CIs) for all metrics were generated through bootstrap simulation to generate 1000 training and test set combinations and thus also 1000 model statistics. This method allows for empiric evaluation of the variability in model predictive power to increase the transparency of model efficacy. We also performed explainability studies for the structured models using Shapley values, also known as SHapley Additive exPlanations values.[30,31] This technique calculates the importance of the features based on the magnitude of feature attributions, using a game theory approach. Shapley values represent the marginal contribution to the model predictions for each feature, calculated over all combinations of subsets of features. We estimated the SHapley Additive exPlanations values on the XGBoost model on the test set.

Results

Study Population

Population characteristics for the entire study cohort of 64 735 patients are summarized in Table 4. Patients who were diagnosed with glaucoma represented 11.22% (N = 7268) of the cohort. The overall mean age was 63.01 years (standard deviation, 15.50). The majority of the cohort was female (61.66%, N = 39 913). The majority of the cohort was non-Hispanic White (60.17%, N = 38 984). Non-Hispanic Black participants formed 16.67% of the cohort (N = 10 794), and Hispanic participants formed 16.26% of the cohort (N = 10 525).

Model Performance

Receiver−operator characteristic curves and precision−recall curves for all models on the entire cohort are shown in Figure 4 evaluated on the test set. The 1D-CNN model achieved the highest performance with an AUROC score of 0.863 (95% CI, 0.862−0.864). XGBoost achieved the second-highest performance with an AUROC score of 0.828 (95% CI, 0.827−0.829). Classification metrics on the overall cohort are summarized in Table 5, with individual classification thresholds tuned on a validation set to maximize F1 score. The 1D-CNN outperformed all models across all metrics except for recall, where logistic regression had the highest performance. The 1D-CNN had the best F1 score of 0.565 (95% CI, 0.561−0.568). In terms of balanced accuracy and precision, the 1D-CNN also performed the best with a score of 0.749 (95% CI, 0.747−0.755) and 0.587 (95% CI, 0.574−0.600), respectively.

We also investigated model performance by cohort race/ethnicity subpopulations. Figure 5 shows the receiver operating characteristic and precision−recall curves for the 1D-CNN stratified by race/ethnicity. The model achieved the highest performance in the non-Hispanic White subpopulation with an AUROC of 0.869 (95% CI, 0.868−0.870). The model had the lowest performance among the non-Hispanic Asian subpopulations with an AUROC of 0.825 (95% CI, 0.820−0.830). Classification metrics stratified by race/ethnicity for the 1D-CNN are shown in Table 6. The model had the best AUROC for the non-Hispanic White population, the model had the best performance for the non-Hispanic Black population for the performance metrics of balanced accuracy at 0.758 (0.755−0.764) and F1 score at 0.593 (0.587−0.599).

Explainability

We performed explainability analyses to determine which input features contributed most to the model predictions, calculating Shapley values for the XGBoost model (Fig 6). The most important features included various demographic and lab results including age, Black race, White race, BMI, hemoglobin A1c, and specific diagnoses. Some of these features are well-known risk factors for glaucoma, such as age and race.

Discussion

In this large nationwide multicenter study, we developed AI prediction algorithms to identify patients who have a high probability of glaucoma, using nonophthalmic EHR diagnoses, medications, laboratory values, and demographics. Despite not using ophthalmic data to predict glaucoma risk, the best-performing model achieved an AUROC of 0.863 (95% CI, 0.862−0.864) for identifying patients with glaucoma. Explainability studies demonstrated that important features included those traditionally known to be risk factors for glaucoma, as well as other systemic features. By using this large and diverse nationwide research platform, we found that our novel approach combining stacked autoencoders and a 1D-CNN architecture outperformed other model architectures in predicting patients at high risk for glaucoma.

There has been some previous work to develop models and tools that use nonimaging data to identify individuals at high risk for certain diagnoses such as glaucoma,[20−22] diabetic retinopathy,[32,33] and other medical diseases.[34−37] Most of these studies primarily rely upon ophthalmic clinical examination findings to risk-stratify patients, thus generally requiring ophthalmic equipment and ophthalmic examination by qualified personnel. For example, Laroche

Table 4. Demographic Population and Lab Result Characteristics

	Glaucoma Patients		Nonglaucoma Patients		Total	
	N = 7268		*N = 57 467*		*N = 64 735*	
Characteristics	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Age	70.80	11.73	62.02	15.64	63.01	15.50
	N	**%**	**N**	**%**	**N**	**%**
Female	4045	55.65%	35 868	62.41%	39 913	61.66%
Race						
White	3917	53.89%	35 031	60.96%	38 948	60.17%
Black	1676	23.06%	9118	15.87%	10 794	16.67%
Asian	206	2.83%	1327	2.31%	1533	2.37%
American Indian or Hawaiian	7	0.10%	50	0.09%	57	0.09%
Middle Eastern or North African	43	0.59%	346	0.60%	389	0.60%
Other	1419	19.52%	11 595	20.18%	13 014	20.10%
Ethnicity						
Hispanic	1092	15.02%	9433	16.41%	10 525	16.26%
Non-Hispanic	5847	80.45%	45 684	79.50%	51 531	79.60%
Other	329	4.53%	2350	4.09%	2679	4.14%
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
BMI	29.67	8.49	30.09	8.05	30.06	8.08
Diastolic BP	75.24	10.6	75.51	10.84	75.67	10.82
Systolic BP	128.53	17.63	127.37	17.77	127.48	17.76
Heart rate	74.56	13.37	75.08	13.85	75.03	13.81
Hemoglobin A1c	6.46	1.56	6.13	1.96	6.16	1.93
Thyroid-stimulating hormone	2.17	3.99	2.25	5.03	2.24	4.95
Lipid panel						
Total cholesterol	181.26	42.01	175.86	43.36	176.43	43.26
LDL cholesterol	101.23	34.8	96.67	35.59	97.12	35.54
HDL cholesterol	54.41	17.94	55.05	17.53	54.98	17.57
Non-HDL cholesterol	124.44	39.6	121.7	39.68	121.88	39.68
Triglyceride	126.16	79.38	124.67	77.85	124.82	78.01
Complete blood count panel						
Red blood cells	4.41	0.65	4.45	3.75	4.45	3.56
White blood cells	7.08	5.8	7.24	3.91	7.22	4.14
Hemoglobin	125.38	32.39	117.91	41.31	118.69	40.54
Hematocrit	39.64	4.72	39.65	5.02	39.65	4.99
Platelets	241.2	74.63	248.27	77.67	247.59	77.41
Comprehensive metabolic panel						
Sodium	139.45	2.77	139.09	2.96	139.13	2.95
Potassium	4.21	0.79	4.2	0.44	4.2	0.49
Chloride	103.47	3.44	103.44	3.47	103.44	3.47
Carbon dioxide	26.93	3.02	26.36	3.22	26.42	3.2
Albumin	39.05	8.53	36.18	12.62	36.47	12.29
Alkaline phosphatase	79.28	35.0	82.31	44.77	82.0	43.87
Bilirubin	0.6	0.38	0.59	1.46	0.59	1.39
Aspartate transaminase	24.85	17.84	24.55	23.6	24.58	23.07
Alanine transaminase	25.1	19.33	23.87	26.5	24.0	25.85
Blood urea nitrogen	17.1	8.56	17.23	9.6	17.21	9.49
Total protein	7.08	0.67	7.06	1.78	7.07	1.7
Calcium	9.35	0.52	9.34	0.54	9.34	0.54
Creatinine	1.07	3.56	1.03	1.21	1.03	1.62
Glucose	113.43	47.5	112.36	48.5	112.47	48.4
pH of urine	6.03	0.85	6.07	1.65	6.07	1.59

BMI = body mass index; BP = blood pressure; HDL = high-density lipoprotein; LDL = low-density lipoprotein; SD = standard deviation.

et al[22] developed a minimally invasive, low-cost calculator method to predict the risk of glaucoma using age, intraocular pressure, and central corneal thickness. Their glaucoma calculator was developed on a cohort of 104 normal and glaucomatous eyes of patients from a single clinic in New York and was able to discriminate between glaucoma patients and controls in this cohort with an AUROC of 0.81.[22] Our models do not include any ophthalmic clinical features at all, which allows them to be used in nonophthalmic settings or on patients who have had no

Table 5. Model Performance Metrics

	AUC	Accuracy	Balanced Accuracy	Sensitivity/Recall	Precision	F1
Logistic regression (L2)	0.686 (0.685, 0.687)	0.63 (0.627, 0.643)	0.629, (0.622, 0.635)	**0.628 (0.611, 0.633)**	0.181 (0.165, 0.187)	0.276 (0.273, 0.278)
XGBoost	0.828 (0.827, 0.829)	0.889 (0.888, 0.892)	0.729, (0.723, 0.733)	0.521 (0.518, 0.525)	0.507 (0.503, 0.511)	0.514 (0.513, 0.517)
FCN	0.796 (0.795, 0.797)	0.891 (0.888, 0.896)	0.662, (0.588, 0.666)	0.366 (0.359, 0.375)	0.527 (0.514, 0.54)	0.43 (0.427, 0.433)
1D-CNN with stacked autoencoder	**0.863 (0.862, 0.864)**	**0.905 (0.901, 0.906)**	**0.749, (0.747, 0.755)**	0.549 (0.538, 0.56)	**0.587 (0.574, 0.6)**	**0.565 (0.561, 0.568)**

1D-CNN = 1-dimensional convolutional neural network; AUC = area under the receiver operating curve; FCN = fully connected network.
Bolded values indicate best performance across all models.

Table 6. Model Performance Metrics on 1D-CNN Stratified by Race/Ethnicity

	AUC	Accuracy	Balanced Accuracy	Sensitivity/Recall	Precision	F1
Hispanic	0.861 (0.860, 0.862)	0.91 (0.904, 0.913)	0.75, (0.744, 0.756)	0.549 (0.541, 0.555)	0.573 (0.554, 0.592)	0.557 (0.549, 0.565)
Non-Hispanic White	0.869 (0.868, 0.870)	0.913 (0.91, 0.916)	0.749, (0.744, 0.754)	0.544 (0.53, 0.558)	0.584 (0.565, 0.603)	0.559 (0.556, 0.565)
Non-Hispanic Asian	0.825 (0.820, 0.830)	0.878 (0.869, 0.888)	0.745, (0.733, 0.758)	0.568 (0.540, 0.596)	0.538 (0.501, 0.576)	0.542 (0.543, 0.564)
Non-Hispanic Black	0.851 (0.850, 0.852)	0.873 (0.868, 0.877)	0.758, (0.755, 0.764)	0.592 (0.576, 0.699)	0.599 (0.584, 0.615)	0.593 (0.587, 0.599)
Non-Hispanic Other	0.843 (0.841, 0.845)	0.897 (0.892, 0.903)	0.757, (0.751, 0.763)	0.576 (0.555, 0.597)	0.565 (0.563, 0.589)	0.564 (0.554, 0.572)

1D-CNN = 1-dimensional convolutional neural network; AUC = area under the receiver operating curve.

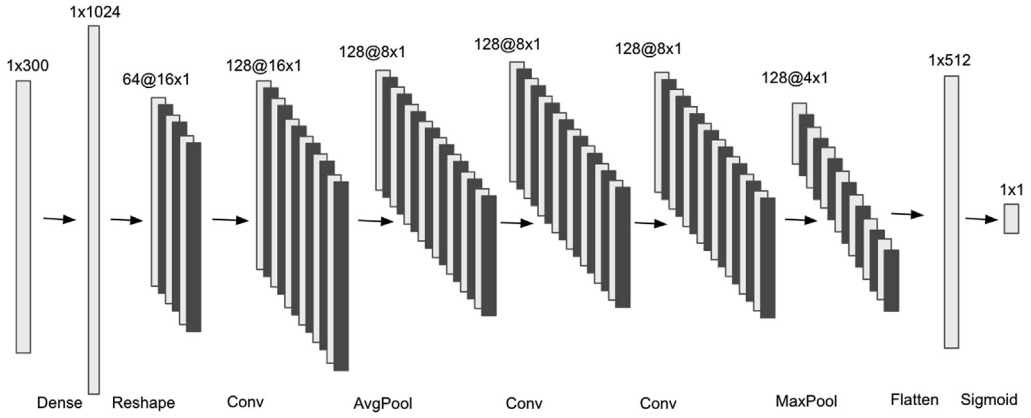

Figure 2. 1D-CNN architecture. Diagram depicts the architecture of our 1D-CNN model. Input is first fed into an FC layer that increases the feature space to create spatial locality and provide enough pixels for the subsequent convolutional layers. Data are then passed through a series of convolutional layers, a max pooling layer, flattened, and finally passed through a sigmoid activation output layer. 1D-CNN = 1-dimensional convolutional neural network; FC = fully connected.

previous eye care at all. In addition, our model was developed on a large multicenter cohort of diverse patients from across the United States and was trained and evaluated on separate populations, such that performance metrics can be reported with less danger of overfitting. A previously developed diabetic retinopathy risk assessment tool (DRRisk) has used a similar approach utilizing only systemic data for their calculator.[32,38] This tool is based on a deep learning model developed on >40 000 patients with diabetes seen at the Los Angeles County Department of Health Services and utilizes 14 features, including systolic blood pressure, hemoglobin A1c, hemoglobin, sex, ethnicity, diastolic blood pressure, age, and others.[32] On a test set of 9300 type 1 and type 2 diabetes patients, their model achieved an AUROC of 0.8.[32] As diabetes is a systemic disease, there are many known systemic markers of poor control that might be reasonably predictive of diabetic retinopathy. Although glaucoma is a different type of disease without as many definitively known systemic risk

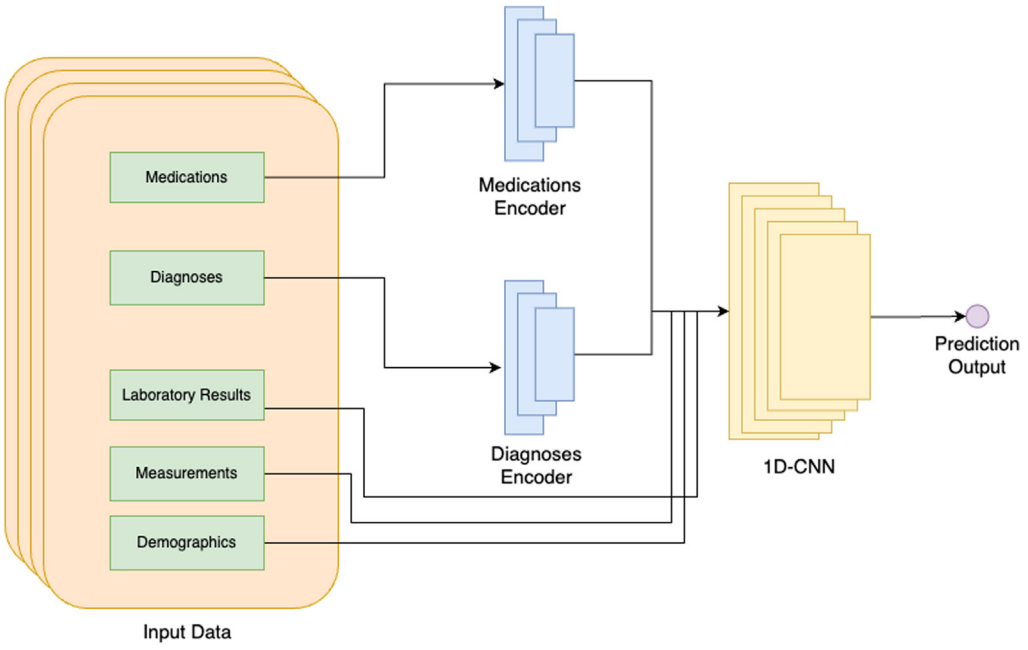

Figure 3. Schematic of prediction pipeline. Diagram illustrates how data flow through the models to output a prediction of whether or not the participant is at high risk for glaucoma. First, medications and diagnoses data are passed and transformed into a reduced-dimensional representation via their respective encoders. Output from both encoders is concatenated with the rest of the input data (i.e., laboratory results, measurements, and demographics) and passed through the 1D-CNN, which outputs a prediction label. 1D-CNN = 1-dimensional convolutional neural network.

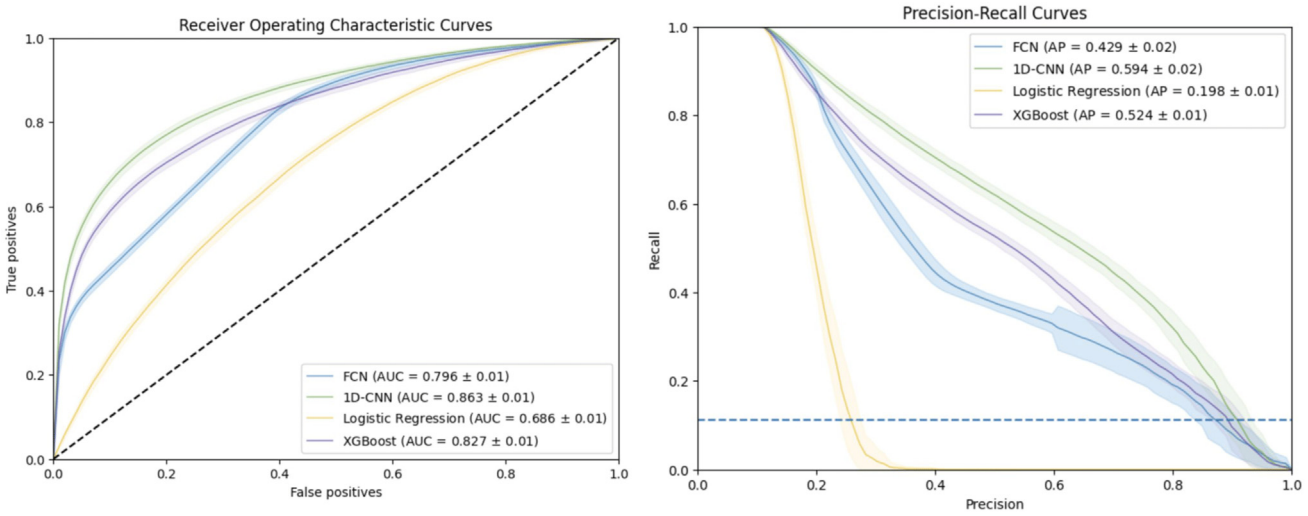

Figure 4. Receiver operating characteristic and precision–recall curves. This figure shows receiver operating characteristic curves (left) and precision–recall curves (right) for all models, evaluated on the entire test set. Curves are shown with 95% confidence interval bands. 1D-CNN = 1-dimensional convolutional neural network; AP = area under the precision-recall curve; AUC = area under the receiver operating curve; FCN = fully connected network.

factors, nevertheless, our EHR model using only systemic features performed remarkably well.

The impact of our approach to glaucoma prescreening using EHR data lies in the potential to enrich the screening population with patients who have a higher probability of glaucoma. Using only routinely collected health data from EHR, such patients identified as highly probable for glaucoma can then be referred for more specialized ophthalmic screening with dedicated imaging, for example. Thus, the time and financial commitment required for dedicated ophthalmic imaging-based screening could be reserved for those with the highest probability of glaucoma, rather than deployed for the general population. General population screening studies for glaucoma have been known to result in a screen-positive rate (or glaucoma prevalence) of approximately 5% at most, highlighting the difficult challenge of general screening for this disease.[13,39–41] For example, Proyecto VER showed a glaucoma prevalence of 1.97% (95% CI, 1.58%–2.36%) in a population-based sample of Hispanic adults older than 40 years.[13] The Rotterdam Study had an overall prevalence of 1.10% (95% CI, 1.09–1.11).[39] It would be challenging to conduct a true randomized controlled trial for glaucoma screening to provide incontrovertible proof of the benefits of early detection of glaucoma when the screen-positive rate is so small, even though we know that treatment of glaucoma prevents vision loss.[42] Our model had a precision (positive predictive value) of up to 0.531, potentially increasing the screen positive rate

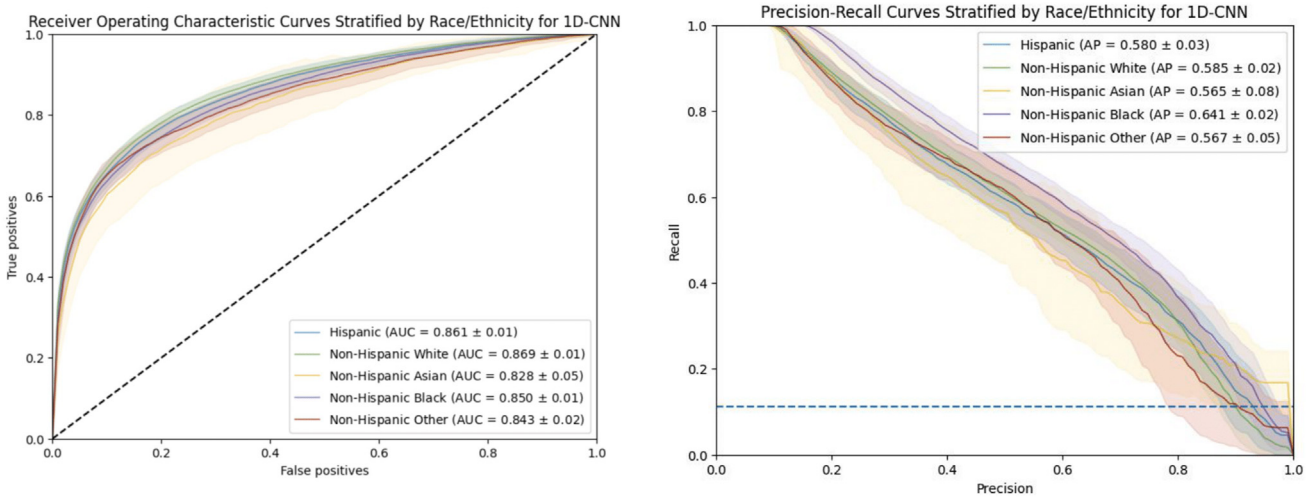

Figure 5. Receiver operating characteristic and precision–recall curves stratified by race/ethnicity for 1D-CNN. This figure shows receiver operating characteristic curves (left) and precision–recall curves (right) for the 1D-CNN model, evaluated on the entire test set and stratified by race/ethnicity. Curves are shown with 95% confidence interval bands. 1D-CNN = 1-dimensional convolutional neural network; AP = area under the precision-recall curve; AUC = area under the receiver operating curve.

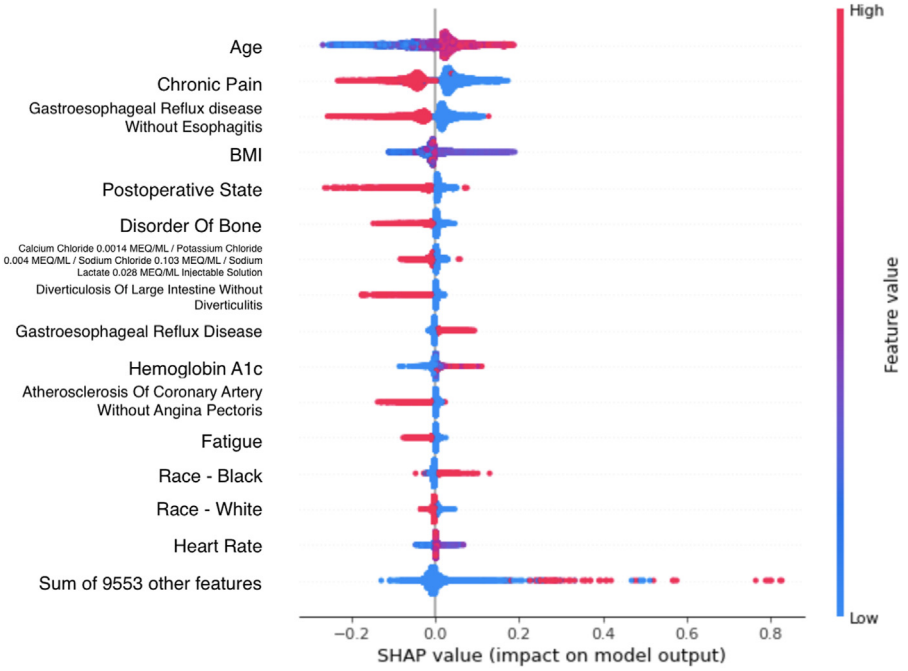

Figure 6. Model explainability with Shapley feature importance. The figure depicts the Shapley value for the topmost important features for predicting whether a patient would be at high risk for glaucoma, using the XGBoost model and calculated across the test set. Positive Shapley values indicate influence toward model prediction of surgery, whereas negative Shapley values indicate influence toward model prediction of no surgery. Points represent individual observations in the test, with the coloring of the points indicating the relative value of that feature for that patient. BMI = body mass index; SHAP = SHapley Additive exPlanations.

by more than 10-fold if only high-probability patients are screened. Such an approach could target resource-intensive screening efforts to those patients most likely to have glaucoma and could also make future randomized clinical trials of targeted glaucoma screening more feasible. We must acknowledge that with such a model, there is a tradeoff between recall (sensitivity) and precision, but an advantage is that for any particular application, the threshold for a "positive" prediction could be adapted to the capacity for imaging-based screening that is available in any specific locale or program. Finally, it is noteworthy that the data used in our study were stored in the Observational Medical Outcomes Partnership Common Data Model,[43] which is an increasingly common standard for representation of health care data from EHRs. The widespread adoption of Observational Medical Outcomes Partnership Common Data Model by numerous health care organizations and data warehouses means that models trained on data using this standard can be more easily tested and implemented in a variety of settings, without the need for additional preprocessing and data harmonization, thus further augmenting their potential usefulness and impact.[44]

An important consideration for glaucoma evaluation in the United States is race and ethnicity. Race is often considered a risk factor for glaucoma: individuals from racial/ethnic minority groups have a higher prevalence of glaucoma compared to non-Hispanic White populations, with prevalence being the highest among the Black population.[11] Researchers have identified novel genetic loci, which have been associated with primary open-angle glaucoma (POAG) in the African Descent population.[45] Researchers have also suggested that there are different single-nucleotide polymorphisms associated with POAG in European Descent and African Descent populations.[45] Previous studies of the POAG susceptibility loci previously associated with European Descent populations have also been associated with Asian populations.[45] Black populations have been shown to have worse glaucoma at diagnosis and worse visual field progression than White patients.[10,46] In addition to genetic influences, there are likely also health disparities as a result of social determinants of health at play.[47] Therefore, when developing an AI model to detect glaucoma in the United States, it is important to investigate model performance in subpopulations of patients, with the goal of avoiding an algorithm that underperforms in minority patients, exacerbating preexisting health disparities. In our study, we did find small but statistically significant differences in the AUROC for our model when stratified by race/ethnicity. For non-Hispanic White patients, the AUROC was 0.869 (95% CI, 0.868−0.870), compared to non-Hispanic Black patients at 0.851 (95% CI, 0.850−0.852) and Hispanic patients at 0.861 (95% CI, 0.860−0.862). This is important to note since such differences can give rise to bias/discrimination during model implementation; therefore, analyzing the robustness of model performance in sensitive subgroups such as by race/ethnicity is important for clinical applications. A comprehensive evaluation of model fairness using additional measures of bias such as

equalized odds or calibration would shed further light on the more nuanced implications of differences in model performance by race/ethnicity. Future research can also investigate model training techniques aimed at mitigating such biases in performance to achieve more fair outcomes.

Another strength of this study is the investigation of model explainability. The features that our model has identified as important are known risk factors for glaucoma, while some are conflicting. Age and Black race are known to be accepted risk factors for glaucoma because older patients are more likely to have glaucoma progression than younger patients at similar intraocular pressure and individuals of African descent are known to be a predictor as there is higher prevalence, earlier presentation, and faster progression of POAG.[11,20] Body mass index was also identified as an important feature for our model. Recent large studies in cohorts in Australia, the United Kingdom, Canada, and South Korea reported significant associations between lower BMI and glaucoma,[48,49] which is consistent with our explainability analyses showing that lower BMI was associated with positive Shapley values, predisposing toward a prediction for glaucoma. While the etiology of the observed relationship is largely unknown, the studies cite potential reasons such as neuroprotective effects of hormones linked to higher BMI, altered cerebrospinal pressure dynamics, or confounding factors like age.[48,49] Our explainability analyses also identified elevated hemoglobin A1c values as a risk factor for glaucoma. Previous literature on the relationship between hemoglobin A1c and glaucoma is complex, as the relationship has many contributing factors including presence and severity of diabetes and types of diabetic therapies used. Some studies have concluded that there is no association between hemoglobin A1c and glaucoma,[50,51] while at the same time other studies have shown that diabetes is a risk factor for POAG.[52] Finally, heart rate was an important feature of our model, which could be indicative of the importance of cardiovascular risk factors and cardiovascular disease in glaucoma, as shown in several previous studies.[53–56] It is important to note that explainability analyses of machine learning models are intended only as "sanity checks" for models to better understand what types of features the model may be relying upon. Explainability studies are not a substitute for traditional causal inference studies; hypothesis-driven inference studies are still needed to rigorously investigate the relationship between individual risk factors and glaucoma. Nevertheless, it is reassuring that so many important features in our model are those already previously shown to have a potential relationship with glaucoma.

Several additional limitations of this study should be acknowledged. Participants in this study were those who had $\geq$2 eye-related diagnoses, such that they would have visited an eye doctor to have an eye examination and either were or were not diagnosed with glaucoma. We could not include all participants in the All of Us in this study, as there would certainly be patients with undiagnosed glaucoma among those without any previous evidence of eye care. Model performance in a truly general population may be different, and future work should validate this model in a true screening situation where all patients regardless of prior eye care are assessed for glaucoma. All of Us does not contain ophthalmic clinical data, imaging, and testing with which to ascertain a glaucoma diagnosis. Hence, we relied upon the billing codes for glaucoma diagnosis, which are subject to miscoding errors. It is possible that some participants with a glaucoma diagnosis did not actually have glaucoma (i.e., were glaucoma suspects). We intentionally excluded participants who were diagnosed only with borderline or suspect glaucoma to help ensure the model would learn to identify glaucoma risk based on participants who clearly did or did not have glaucoma. However, due to miscoding, it is possible that some glaucoma suspects remained in the population. Additional future validation studies utilizing data registries that contain dedicated ophthalmic clinical data, such as the Sight Outcomes Research Collaborative, or institutional repositories that contain ophthalmic clinical imaging, would be additional next steps in validating this model.

In conclusion, we have shown that machine and deep learning models can analyze the wealth of systemic data stored in EHRs to identify patients with a high probability of glaucoma, without the use of any ophthalmic imaging or clinical data. Our 1D-CNN deep learning architectures performed especially well compared with traditional machine learning approaches. There were small but significant differences in performance between race/ethnicity subgroups, illustrating the importance of developing and investigating AI algorithm performance in a diverse cohort of patients. Explainability studies showed that many important input features were clinically reasonable. Algorithms such as ours, which identify patients at high risk of glaucoma from only systemic EHR data, could eventually enable resource-intensive glaucoma screening to be targeted to those patients with the highest risk.

Footnotes and Disclosures

Originally received: June 19, 2024.
Final revision: October 31, 2024.
Accepted: December 2, 2024.
Available online: December 6, 2024. Manuscript no. XOPS-D-24-00197.

Department of Ophthalmology, Byers Eye Institute, Stanford University, Palo Alto, California.

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

The authors have no proprietary or commercial interest in any materials discussed in this article.

Financially supported by the National Eye Institute K23EY03263501 (S.Y.W.); Career Development Award from the Research to Prevent Blindness (S.Y.W.); unrestricted departmental grant from the Research to Prevent Blindness (all authors); and departmental grant from the National Eye Institute P30-EY026877 (all authors). The funders had no role in the design and conduct of the study; collection, management, analysis, and

interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Support for Open Access publication was provided by the Department of Ophthalmology, Stanford University.

HUMAN SUBJECTS: No human subjects were included in this study. This study using deidentified data was exempt from Stanford University Institutional Review Board approval and adheres to the tenets of the Declaration of Helsinki.

No animal subjects were used in this study.

Author Contributions:

Conception and design: Wang, Ravindranath

Data collection: Ravindranath

Analysis and interpretation: Wang, Ravindranath

Obtained funding: Wang

Overall responsibility: Wang, Ravindranath

Abbreviations and Acronyms:
1D-CNN = 1-dimensional convolutional neural network; **AI** = artificial intelligence; **AUROC** = area under the receiver operating characteristic curve; **BMI** = body mass index; **CI** = confidence interval; **EHR** = electronic health record; **POAG** = primary open-angle glaucoma.

Keywords:
Glaucoma screening, Machine learning, Deep learning, Electronic health records.

Correspondence:
Sophia Y. Wang, MD, MS, Byers Eye Institute, Stanford Hospital, 2370 Watson Ct, Palo Alto, CA 94303. E-mail: sywang@stanford.edu.

References

1. Resnikoff S, Pascolini D, Etya'ale D, et al. Global data on visual impairment in the year 2002. *Bull World Health Organ*. 2004;82:844−851.
2. Kingman S. Glaucoma is second leading cause of blindness globally. *Bull World Health Organ*. 2004;82:887−888.
3. Zhang N, Wang J, Li Y, Jiang B. Prevalence of primary open angle glaucoma in the last 20 years: a meta-analysis and systematic review. *Sci Rep*. 2021;11:13762.
4. Davuluru SS, Jess AT, Kim JSB, et al. Identifying, understanding, and addressing disparities in glaucoma care in the United States. *Transl Vis Sci Technol*. 2023;12:18.
5. Gupta P, Zhao D, Guallar E, et al. Prevalence of glaucoma in the United States: the 2005-2008 national health and nutrition examination survey. *Invest Ophthalmol Vis Sci*. 2016;57:2905−2913.
6. Shaikh Y, Yu F, Coleman AL. Burden of undetected and untreated glaucoma in the United States. *Am J Ophthalmol*. 2014;158:1121−1129.e1. Elsevier.
7. Rodriguez J, Sanchez R, Munoz B, et al. Causes of blindness and visual impairment in a population-based sample of U.S. Hispanics. *Ophthalmology*. 2002;109:737−743. Elsevier.
8. Giangiacomo A, Coleman AL. The epidemiology of glaucoma. In: Grehn F, Stamper R, eds. *Glaucoma*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2009:13−21.
9. Kang JH, Wang M, Frueh L, et al. Cohort study of race/ethnicity and incident primary open-angle glaucoma characterized by autonomously determined visual field loss patterns. *Transl Vis Sci Technol*. 2022;11:21.
10. Halawa OA, Jin Q, Pasquale LR, et al. Race and ethnicity differences in disease severity and visual field progression among glaucoma patients. *Am J Ophthalmol*. 2022;242:69−76.
11. Allison K, Patel DG, Greene L. Racial and ethnic disparities in primary open-angle glaucoma clinical trials: a systematic review and meta-analysis. *JAMA Netw Open*. 2021;4:e218348.
12. Tielsch JM, Sommer A, Katz J, et al. Racial variations in the prevalence of primary open-angle glaucoma. The Baltimore Eye Survey. *JAMA*. 1991;266:369−374.
13. Quigley HA, West SK, Rodriguez J, et al. The prevalence of glaucoma in a population-based study of Hispanic subjects: Proyecto VER. *Arch Ophthalmol*. 2001;119:1819−1826.
14. US Preventive Services Task Force, Mangione CM, Barry MJ, Nicholson WK. Screening for primary open-angle glaucoma: US preventive Services Task Force recommendation statement. *JAMA*. 2022;327:1992−1997.
15. Karvonen E, Stoor K, Luodonpää M, et al. Combined structure-function analysis in glaucoma screening. *Br J Ophthalmol*. 2022;106:1689−1695.
16. Tatemichi M, Nakano T, Tanaka K, et al; Glaucoma Screening Project (GSP) Study Group. Performance of glaucoma mass screening with only a visual field test using frequency-doubling technology perimetry. *Am J Ophthalmol*. 2002;134:529−537.
17. Wilson MR, Khanna S. The value of different screening techniques for glaucoma. *Curr Opin Ophthalmol*. 1994;5:69−75.
18. Li G, Fansi AK, Harasymowycz P. Screening for glaucoma using GDx-VCC in a population with ≥1 risk factors. *Can J Ophthalmol*. 2013;48:279−285.
19. de Vries MM, Stoutenbeek R, Müskens RPHM, Jansonius NM. Glaucoma screening during regular optician visits: the feasibility and specificity of screening in real life. *Acta Ophthalmol*. 2012;90:115−121.
20. Guedes RAP, Guedes VMP, Chaoubah A. Focusing on patients at high-risk for glaucoma in Brazil: a pilot study. *J Fr Ophtalmol*. 2009;32:640−645.
21. Paudyal I, Yadav R, Parajuli A, et al. Screening of accompanying first degree relatives of patients with primary open angle glaucoma. *Nepal J Ophthalmol*. 2022;14:4−9.
22. Laroche D, Rickford K, Mike EV, et al. A novel, low-cost glaucoma calculator to identify glaucoma patients and stratify management. *J Ophthalmol*. 2022;2022:5288726.
23. All of Us Research Program Investigators, Denny JC, Rutter JL, Goldstein DB. The "All of Us" research program. *N Engl J Med*. 2019;381:668−676.
24. Mapes BM, Foster CS, Kusnoor SV, et al. All of Us Research Program. Diversity and inclusion for the All of Us research program: a scoping review. *PLoS One*. 2020;15:e0234962.
25. OMOP CDM v5.3. https://ohdsi.github.io/CommonDataModel/cdm53.html. Accessed January 18, 2024.
26. Murti DMP, Pujianto U, Wibawa AP, Akbar MI. *"K-Nearest Neighbor (K-NN) based Missing Data Imputation," 2019 5th International Conference on Science in Information Technology (ICSITech)*. 2019:83−88. Yogyakarta, Indonesia.
27. "Learn," scikit. https://scikit-learn.org/1.2/. Accessed February 2, 2024.
28. Tensorflow TensorFlow https://www.tensorflow.org/. Accessed February 2, 2024.
29. "Autoencoders," Unsupervised Feature Learning and Deep Learning Tutorial. http://ufldl.stanford.edu/tutorial/unsupervised/Autoencoders/. Accessed December 11, 2023.

30. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, et al., eds. *Advances in Neural Information Processing Systems*. 30. Curran Associates, Inc.; 2017.
31. Lundberg S. shap. Github. https://github.com/slundberg/shap. Accessed December 23, 2023.
32. Ogunyemi OI, Gandhi M, Lee M, et al. Detecting diabetic retinopathy through machine learning on electronic health record data from an urban, safety net healthcare system. *JAMIA Open*. 2021;4:ooab066.
33. Related retinopathy risk test. Diabetes https://diabetes.org/retinopathy-risk-test. Accessed December 18, 2023.
34. Sheth S, Lee P, Bajaj A, et al. Implementation of a machine-learning algorithm in the electronic health record for targeted screening for familial hypercholesterolemia: a quality improvement study. *Circ Cardiovasc Qual Outcomes*. 2021;14:e007641.
35. Anderson AE, Kerr WT, Thames A, et al. Electronic health record phenotyping improves detection and screening of type 2 diabetes in the general United States population: a cross-sectional, unselected, retrospective study. *J Biomed Inform*. 2016;60:162−168.
36. Shao Y, Zeng QT, Chen KK, et al. Detection of probable dementia cases in undiagnosed patients using structured and unstructured electronic health records. *BMC Med Inf Decis Making*. 2019;19:128.
37. Wang L, Laurentiev J, Yang J, et al. Development and validation of a deep learning model for earlier detection of cognitive decline from clinical notes in electronic health records. *JAMA Netw Open*. 2021;4:e2135174.
38. Gandhi M, Daskivich LP, Ogunyemi OI. DRRisk: a web-based tool to assess the risk of diabetic retinopathy through machine learning on electronic health records. *AMIA Annu Symp Proc*. 2023;2022:452−460.
39. Klein BE, Klein R, Sponsel WE, et al. Prevalence of glaucoma. The beaver dam eye study. *Ophthalmology*. 1992;99:1499−1504.
40. Dielemans I, Vingerling JR, Wolfs RC, et al. The prevalence of primary open-angle glaucoma in a population-based study in The Netherlands. The Rotterdam Study. *Ophthalmology*. 1994;101:1851−1855.
41. Coffey M, Reidy A, Wormald R, et al. Prevalence of glaucoma in the west of Ireland. *Br J Ophthalmol*. 1993;77:17−21.
42. Heijl A, Leske MC, Bengtsson B, et al. Reduction of intraocular pressure and glaucoma progression: results from the early manifest glaucoma trial. *Arch Ophthalmol*. 2002;120:1268−1279.
43. Data standardization. OHDSI https://www.ohdsi.org/data-standardization/. Accessed February 5, 2024.
44. "Observational Health Data Sciences and Informatics," resources:2020_data_network [Observational Health Data Sciences and Informatics]. https://www.ohdsi.org/web/wiki/doku.php?id=resources%3A2020_data_network. Accessed February 5, 2024.
45. Zukerman R, Harris A, Verticchio Vercellin A, et al. Molecular genetics of glaucoma: subtype and ethnicity considerations. *Genes*. 2021;12:55.
46. Elam AR, Andrews C, Musch DC, et al. Large disparities in receipt of glaucoma care between enrollees in medicaid and those with commercial health insurance. *Ophthalmology*. 2017;124:1442−1448.
47. Siegfried CJ, Shui YB. Racial disparities in glaucoma: from epidemiology to pathophysiology. *Mo Med*. 2022;119:49−54.
48. Marshall H, Berry EC, Torres SD, et al. Association between body mass index and primary open angle glaucoma in three cohorts. *Am J Ophthalmol*. 2023;245:126−133.
49. Lin SC, Pasquale LR, Singh K, Lin SC. The association between body mass index and open-angle glaucoma in a South Korean population-based sample. *J Glaucoma*. 2018;27:239−245.
50. Johnson NA, Jammal AA, Berchuck SI, Medeiros FA. Effect of diabetes control on rates of structural and functional loss in patients with glaucoma. *Ophthalmol Glaucoma*. 2021;4:216−223.
51. Johnson N, Adad Jammal A, Medeiros FA, et al. Effect of diabetes control on rates of visual field loss in patients with glaucoma. *Invest Ophthalmol Vis Sci*. 2020;61:3891.
52. Zhao YX, Chen XW. Diabetes and risk of glaucoma: systematic review and a Meta-analysis of prospective cohort studies. *Int J Ophthalmol*. 2017;10:1430−1435.
53. Marshall H, Mullany S, Qassim A, et al. Cardiovascular disease predicts structural and functional progression in early glaucoma. *Ophthalmology*. 2021;128:58−69.
54. Kurysheva NI, Shlapak VN, Ryabova TY. Heart rate variability in normal tension glaucoma: a case-control study. *Medicine (Baltim)*. 2018;97:e9744.
55. Choi JA, Lee SN, Jung SH, et al. Association of glaucoma and lifestyle with incident cardiovascular disease: a longitudinal prospective study from UK Biobank. *Sci Rep*. 2023;13:2712.
56. Lee EB, Hu W, Singh K, Wang SY. The association among blood pressure, blood pressure medications, and glaucoma in a nationwide electronic health records database. *Ophthalmology*. 2022;129:276−284.